Sobre a resolução de correferência

Tatiane de Moraes Coreixas[1], Renata Vieira[1]

[1]Faculdade de Informática – Pontifícia Universidade Católica do RS (PUC)

tatiane.coreixas@pucrs.br, renata.vieira@pucrs.br

Abstrat. *Coreference resolution is an important task in many Natural Language Processing applications. This task often depends on the understanding of textual context and world knowledge. These are complex linguistic and extra linguistic questions related to human cognitive abilities, that are not easily reproduced by computer systems. In order to explain the difficulty of the task of coreference resolution, a study of linguistic and technological aspects of this task is developed and described in this work.*

Resumo. *A resolução de correferência é uma tarefa importante em várias aplicações de Processamento da Linguagem Natural. Essa tarefa é muitas vezes dependente da compreensão do contexto textual e conhecimento de mundo. Estas são questões lingüísticas e extralingüísticas complexas relacionadas às habilidades cognitivas humanas e que não são facilmente reproduzidas por sistemas computacionais. Com o objetivo de explicar a dificuldade da tarefa de resolução de correferência, um estudo sobre aspectos lingüísticos e tecnológicos dessa tarefa é elaborado e descrito nesse trabalho.*

Palavras-chave: correferência, processamento da língua natural

1. Introdução

Com o surgimento da tecnologia, o acesso à informação ficou mais ágil e fácil. Essa facilidade proporcionou uma sobrecarga de informação. Devido a esse cenário, as ferramentas computacionais precisam aprimorar-se e adequar-se às novas necessidades.

Uma destas necessidades é a compreensão da Língua Natural (LN) por sistemas computacionais. A área que se ocupa destes estudos é o Processamento da Linguagem Natural (PLN), um ramo da Inteligência Artificial que estuda os problemas relacionados com a língua. Seu objetivo é desenvolver recursos tecnológicos para auxiliar nos variados processos de comunicação em língua natural.

A área de PLN teve início no final da década de 40, com a tradução automática de documentos. Em uma primeira fase de desenvolvimento a área preocupou-se principalmente com o estudo das frases isoladas, sem o foco na identificação do contexto.

Com os avanços nessa área de pesquisa, surgiram novas áreas de interesse, como a recuperação de informação e a sumarização automática, dentre outras. Com isso, a necessidade de realizar a compreensão global do contexto tornou-se fundamental para alcançar os novos objetivos. É nesse ponto que está uma das grandes dificuldades do PLN, pois essa é uma habilidade cognitiva bem desenvolvida no ser humano e difícil de reproduzir automaticamente.

Para isso, surge a interação com áreas como o estudo do discurso. Além de analisar frases, os sistemas devem envolver a construção do significado projetado no decorrer do texto.

> "Há diversas razões para o fato de a compreensão de linguagem natural ter se mostrado uma tarefa difícil para a comunidade de IA. Entre as razões mais importantes estão a grande quantidade de conhecimento, de habilidades e de experiências necessárias para dar suporte ao uso da linguagem." [LUG04]

Os estudos lingüísticos abrangem diversos focos de estudo. A área de fonética e fonologia estuda as formas sonoras da língua. Em PLN esse estudo está envolvido com a capacidade de desenvolver sistemas relacionados à comunicação oral, que possam interpretar e sintetizar a linguagem falada. A morfologia estuda principalmente questões relativas às palavras, enquanto a sintaxe preocupa-se com o agrupamento de palavras na formação das frases, com o objetivo de estudar e descrever a organização dos elementos de uma frase e suas regras de formação. A semântica estuda o significado dos elementos lingüísticos, isto é, sua relação com entidades do mundo. A pragmática e o discurso estudam o uso da língua e sua relação com falante e o contexto, estes estudos consideram o texto como unidade, abordando fatores de compreensão textual, como coesão e coerência.

Os estudos lingüísticos são importantes para a resolução automática de correferência (identificar as referências às entidades e seus relacionamentos no texto), tendo em vista que essa é uma tarefa que abrange o conhecimento lingüístico em vários níveis, incluindo o conhecimento do contexto. A resolução de correferência é uma tarefa fundamental para determinadas áreas do PLN, como a sumarização automática e a extração de informação. Apesar dos avanços obtidos nestas áreas, ainda há muitos problemas em aberto.

2. Resolução de Correferência

A tarefa de resolução de correferência é considerada de alta complexidade devido à sua abrangência, considerando os diferentes níveis lingüísticos necessários para seu tratamento. Define-se correferência como a relação entre elementos lingüísticos que se referem a uma mesma entidade de mundo. Sua resolução tem como intuito deixar a relação entre as entidades evidenciada.

Para o entendimento de correferência, é preciso definir anáfora, já que seus conceitos estão relacionados.

Anáfora se define como toda a retomada de idéia já introduzida por uma entidade mencionada anteriormente. Pode-se dizer que quando uma entidade é citada pela primeira vez em um texto, se faz o processo de evocação da entidade. A expressão que realiza o acesso é dita anafórica, e a expressão a quem ela se refere é chamada de antecedente. Sendo que a relação entre essas expressões denomina-se relação de correferência.

No Exemplo 1 é apresentado um exemplo de correferência extraído de Abreu [ABR05]. As expressões correferentes estão destacadas.

"O advogado de Castor de Andrade, Nélio Machado, afirmou que vai aguardar a evolução dos fatos para se pronunciar. O advogado disse desconhecer a existência de documentos que demonstrariam o pagamento de propinas as autoridades policiais".

Exemplo 1. Exemplo de correferência entre entidades

Percebe-se que a expressão "o advogado" refere-se a "o advogado de Castor de Andrade". Esta relação entre as expressões é que se denomina de correferência.

Os conceitos de correferência e anáfora são semelhantes. Expressões correferentes fazem referência à mesma entidade, já expressões anafóricas retomam uma referência ou ativam um novo referente. Geralmente uma expressão correferente é anafórica, mas nem sempre uma expressão anafórica é correferente. Observe o Exemplo 2 (retirado de [ABR05]).

"O Eurocenter oferece cursos de Japonês na bela cidade de Kanazawa. Os cursos têm quatro semanas de duração. As aulas do nível avançado incluem refeições típicas e passeios a pontos turísticos".

Exemplo 2. Exemplo de expressões anafóricas

De acordo com o exemplo 2, a expressão "Os cursos" retoma uma expressão já citada no discurso, "cursos de Japonês", sendo que essas duas expressões fazem menção à mesma entidade, são expressões correferenciais e anafóricas. Já a expressão "As aulas do nível avançado" não é correferente a nenhum termo, mas apresenta significado na expressão "cursos de Japonês", sendo uma expressão anafórica, mas não correferente.

De acordo com estudos realizados por Vieira [VIE98], o relacionamento da anáfora com seu antecedente podem ser divididos da seguinte forma (todos os exemplos abaixo foram extraídos de [ROS02]):

Anáforas diretas: são antecedidas por uma expressão, sendo que possui o mesmo nome-núcleo e a entidade no discurso a que se referem é a mesma, observe o exemplo abaixo:

"A proposta está no projeto de lei que regulamenta a reforma da Previdência e foi enviado ontem ao Congresso pelo presidente Fernando Henrique Cardoso. Pelo projeto, as contribuintes autônomas passarão a ter direito ao salário-maternidade."

Exemplo 3. Exemplo de anáfora direta

Anáforas Indiretas: são antecedidas por uma expressão que não têm o mesmo nome-núcleo do seu antecedente, mas a entidade no discurso a que se referem é a mesma. Assim, o núcleo pode ser um sinônimo do antecedente.

"O Sindicato e a Associação dos Bancos mantêm a tese de que somente a União pode legislar em questões de ordem financeira. (...) O assessor jurídico das entidades, Flávio do Couto Silva, garante que não existe qualquer dúvida quanto à proteção judicial pelos bancos."

Exemplo 4. Exemplo de anáfora indireta

Anafóricas Associativas: introduzem um referente novo no discurso, mas seu significado está ancorado em um ente e, por isso, não podem ser classificadas como nova no discurso.

"O ministro da Indústria e Comércio da Argentina, Alieto Guadagni, disse ontem que o Mercosul está 'agonizando', mas ainda é possível salvá-lo se os sócios se comprometerem a fortalecê-lo institucionalmente."

Exemplo 5. Exemplo de anáfora associativa

Pode-se dizer que expressões anafóricas diretas e indiretas são correferentes, pois fazem referência à mesma entidade. Enquanto que expressões anafóricas associativas são não correferentes, pois não se referem a mesma entidade.

A anáfora direta pode ser considerada como um substituto do elemento que está sendo retomado por ela [MAR05]. Aspectos gramaticais como concordância de gênero e número, influenciarão na escolha do antecedente referencial, principalmente quando houver mais de um candidato. De acordo com [MAR05] "... a visão clássica da anáfora direta se dá com base na noção de que a anáfora é um processo de reativação de referentes prévios."

Uma nova referência para uma entidade pode introduzir uma nova informação sobre ela, e várias entidades são referenciadas. Assim, existe um conjunto de referências e um conjunto de referentes em potencial. Essas informações podem ser agrupadas formando as cadeias de correferência.

Desta forma, o problema de resolução de correferência pode ser visto como: tendo um conjunto de expressões referenciais de um texto necessita-se identificar quais são os subconjuntos de sintagmas que fazem referência a mesma entidade. As substituições realizadas pelo autor nos processos de correferência textual são importantes para não tornar o texto repetitivo, assim como para manter a coesão textual.

Para resolver as referências automaticamente, faz-se necessário o uso de conhecimento dos variados níveis de processamento lingüístico. É preciso realizar a análise sintática das frases, análise do contexto e semântica, o que pode envolver conhecimento de mundo. Com uma combinação dessas informações, é possível realizar a identificação das expressões correferentes. Porém, capturar um modelo do contexto é ainda um desafio grande para a área.

3. Sistemas e recursos para o processamento de correferência

A seguir, serão apresentados alguns sistemas que realizam o processamento de correferência e recursos necessários para o desenvolvimento e avaliação desses

sistemas.

3.1. Resolução de pronomes anafóricos

O algoritmo apresentado por Chaves [CHA07], apresenta uma adaptação do algoritmo de Mitkov para a resolução de anáforas na língua portuguesa, sendo estas compostas por pronomes pessoais de terceira pessoa, o que determinada que os seus antecedentes devam ser sintagmas nominais.

O algoritmo possui um funcionamento simples e consiste na aplicação de heurísticas para a identificação do antecedente da anáfora, dispensando o uso de conhecimento lingüístico profundo.

As heurísticas estabelecidas estão divididas da seguinte forma: as promocionais (primeiro sintagma nominal da sentença, padrão de colocação, reiteração lexical, paralelismo sintático, sintagma nominal mais próximo e nome próprio); as impeditivas (sintagma nominal indefinido e sintagma nominal preposicionados) e a distância referencial, sendo que esta pode ser tanto promocional ou impeditiva, pois sua pontuação é de acordo com a posição dos candidatos em relação à anáfora. A maior pontuação é fornecida para aqueles que se encontram na mesma sentença da anáfora.

Pode-se observar que há uma combinação de heurísticas morfossintáticas elaboradas e heurísticas que empregam conhecimento preliminar de discurso, tais como proximidade e distância referencial.

3.2. O algoritmo de Lappin e Leass adaptado para o português

O trabalho desenvolvido por [COE05] foi de realizar uma adaptação do algoritmo de Lappin e Leass para a língua portuguesa. Para a realização desse trabalho foi utilizado um corpus anotado com informações morfológicas e sintáticas.

O processamento realizado pelo algoritmo é de identificar os possíveis candidatos de anáfora nas sentenças, criando classes de equivalência para os candidatos e, assim, gerar uma lista com os possíveis candidatos para os pronomes de terceira pessoa, concordando em gênero e número, com base na atribuição de pesos adicionais aos candidatos e, com isso, selecionar o candidato com maior fator de saliência. O fator de desempate, caso ocorra, será o candidato mais próximo ao pronome [COE05].

3.3. Resolução de correferência

O sistema desenvolvido por Souza [SOU07] tem por objetivo automatizar a resolução de correferência para a língua portuguesa, através de uma abordagem baseada em aprendizado de máquina supervisionado. O sistema seleciona as cadeias de correferência de um texto, através da identificação dos pares de expressões anafóricas.

Esses pares são gerados de acordo com heurísticas pré-estabelecidas pelo autor, que são: comparação dos núcleos do par de sintagmas; distância em número de frases entre os dois sintagmas; verificação se o antecedente é pronome; se a anáfora é pronome; se os sintagmas são nomes próprios; concordância de gênero; concordância de número; se os sintagmas são sujeitos; concordância semântica (se possuem etiquetas semânticas idênticas; e mesmo grupo semântico (caso possuam etiquetas semânticas que pertençam ao mesmo grupo).

Aqui também observa um conjunto de informações lingüísticas que remetem a

níveis distintos, tais como sintáticos, semânticos e de discurso.

3.4. O Corpus Summ-It

O Summ-It é um *corpus* composto por cinqüenta textos jornalísticos do caderno de Ciências da Folha de São Paulo retirados do corpus PLN-BR. Os textos foram anotados com informação sintática, de correferência e de estrutura retórica. O Summ-it também conta com sumários, sendo que para cada um de seus textos há sumários e extratos automáticos de vários tipos.

Cada documento do corpus corresponde a um arquivo texto com tamanho entre 1 e 4KB (de 127 a 654 palavras). Dos 5047 *markables* (sintagmas anotados), a maior parte corresponde a sintagmas nominais com nome núcleo (95,15%). Pronomes são constituem apenas 4,82%.

A anotação de correferência do Summ-It fez uso do analisador sintático de textos em português PALAVRAS e da ferramenta de anotação MMAX.

Na anotação de correferência (e anáforas), as unidades de interesse são os sintagmas nominais, considerando, evidentemente, as entidades mencionadas e estruturas simples (por exemplo, de pronomes substantivos), e desconsiderando as estruturas oracionais. Os markables foram gerados de modo semi-automático, isto é, primeiramente, as unidades de interesse foram selecionadas automaticamente a partir da anotação em formato XML do PALAVRAS e, posteriormente, esse resultado foi revisado. O segundo passo foi a identificação das configurações morfossintáticas dos markables. Esse processo, assim como o anterior, ocorreu de modo semi-automático, após a geração automática, o resultado foi revisado manualmente. O terceiro passo, a construção das relações das cadeias de correferência foi realizada manualmente por uma equipe de 12 anotadores, sendo que cada texto foi anotado por dois membros dessa equipe.

Pode-se perceber que o procedimento para realizar a anotação de correferência não é uma tarefa simples. O trabalho manual exigido para esse tipo de tarefa torna custosa a realização de anotação de correferência em uma base constituída por muitos textos. Escassez de recursos e dificuldade na construção dos sistemas são fatores decisivos nesse cenário. Por envolver a interpretação humana, há a necessidade de mais de um revisor para a realização da tarefa, para garantir a confiabilidade no resultado da anotação.

3.5. MUC e ACE

O MUC (*Message Understanding Conference*) [MUC97] foi uma conferência destinada a tratar questões relacionadas com a resolução de correferência para a língua inglesa. Teve início em 1987, sendo que seu último encontro foi em 1998.

Primeiramente, os participantes deveriam desenvolver e treinar seus algoritmos com base em um corpus fornecido pela organização da conferência (corpus MUC). Após um período de tempo (em torno de 6 meses), os algoritmos eram avaliados numa nova base de dados. Os resultados atingidos pelos algoritmos eram comparados com o resultado obtido manualmente.

O ACE (*Automatic Content Extraction*) [ACE08] é uma continuação dos trabalhos iniciados no MUC, e também apresenta a identificação e classificação das

referências anafóricas entre as tarefas propostas para avaliação de sistemas de extração de informação. Sua coleção de textos inclui o inglês, chinês e o árabe. Um corpus é disponibilizado para avaliação.

3.6. HAREM

No âmbito da língua portuguesa, foi criado o HAREM [HAR08], que é uma atividade de avaliação conjunta com o intuito de incentivar as pesquisas de PLN para a língua portuguesa. Em 2008 uma tarefa relacionada ao reconhecimento de relações entre entidades nomeadas foi proposta pela primeira vez. O HAREM disponibiliza um *corpus* com as anotações de referências. Esse *corpus* serve para conferência dos sistemas participantes, e é denominado de coleção dourada.

4. Considerações finais

A tarefa de resolução de correferência apresenta ainda um grande desafio para o PLN. As dificuldades dessa tarefa estão relacionadas com a necessidade de tratar vários níveis lingüísticos, incluindo semântico e discursivo-pragmático que na área de PLN apresentam muitos problemas em aberto. A tentativa de desenvolver um sistema de abrangência geral torna a tarefa ainda mais complexa.

Entre os casos de fácil resolução estão as repetições simples (anáforas diretas). Os principais problemas encontrados por esses sistemas estão relacionados com os tipos de anáfora indireta (incluindo anáforas pronominais) e associativa. Nesses tipos de anáfora, o núcleo das expressões co-referentes não são os mesmos, como ocorre no caso das anáforas diretas. Identificar a relação entre os referentes neste caso depende do conhecimento semântico que se estabelece não só através das relações entre as palavras (como no caso de relações de sinonímia e hiperonímia), mas também é influenciado pelo contexto textual e conhecimento de mundo.

Durante o levantamento das entidades correferentes, geralmente mais de uma expressão é considerada como candidata a entidade. Decidir quais antecedentes são válidos para montar as cadeias de correferência não é uma tarefa simples.

Para o desenvolvimento dessa área de pesquisa a construção de corpus também se faz necessária. Essa é uma atividade que envolve a marcação manual dos textos com informações lingüísticas de vários níveis além da identificação das entidades correferentes. Este trabalho requer, muitas vezes, uma etapa de formação e treinamento de uma equipe, que é um processo demorado e caro. Outra questão a ser considerada é a subjetividade da tarefa. Uma convergência entre as respostas deve ser almejada, demonstrando confiabilidade das anotações produzidas.

Essa base de dados lingüísticos, o *corpus*, é muito importante para a tarefa de resolução da correferência. Os sistemas desenvolvidos utilizam esses textos marcados para realizarem seus respectivos testes, como forma de validação de seus resultados.

Devido à dependência do nível semântico, a resolução da correferência talvez consiga melhores resultados em domínios restritos, através de bases de conhecimento específicas de domínio.

5. Referências bibliográficas

[ACE08] **ACE - Automatic Content Extraction.** Disponível em: http://www.nist.gov/speech/tests/ace/2008/. Acessado em: 19 de junho de 2008.

[ABR05] ABREU, Sandra C. de. **Análise de expressões referenciais em corpus anotado da Língua Portuguesa.** Dissertação de Mestrado, Unisinos, São Leopoldo, 2005.

[CHA07] CHAVES, Amanda R.; RINO, Lucia H. M. **A resolução de pronomes anafóricos do português com base em heurísticas que apontam o antecedente.** Congresso de Pós- Graduação, Universidade Federal de São Carlos, SP, 2007.

[COE05] COELHO, Thiago; CARVALHO, Ariadne. **Uma adaptação do algoritmo de Lappin e Leass para resolução de anáforas em português.** SBC, 2005.

[HAR08] **HAREM: Reconhecimento de entidades mencionadas em português.** Disponível em: http://www.linguateca.pt/HAREM/. Acessado em: 20 de junho de 2008.

[LUG04] LUGER, George F. **Inteligência Artificial: Estruturas e estratégias para a solução de problemas complexos.** Tradução de: Paulo Engel. Porto Alegre, Bookman, 2004.

[MAR05] KOCH, Ingedore G. V (Org.); MARCUSCHI, Luiz Antônio. **Referenciação e discurso.** São Paulo, Contexto, 2005, p. 33-101.

[MUC97] **MUC-7 - Coreference task definition. Proceedings of the Seventh Message Understanding Conference (MUC-7).** San Francisco, CA, 1997.

[ROS02] ROSSI, Daniela, et al. **Resolução de Correferência em texto da língua portuguesa.** SBC, Unisinos, São Leopoldo,

[SOU07] SOUZA, J. G. C. de. **Resolução automática de correferência aplicada à língua portuguesa.** Trabalho de conclusão. Unisinos, São Leopoldo, 2007.

[VIE98] VIEIRA, Renata. **Definite description processing in unrestricted text.** Tese de Doutorado, University of Edinburgh, Edinburgh, 1998.b